



## Section 2


---

# Hadoop Assignment I

Using Hadoop's Core:  
HDFS and MapReduce



# My Notes



```
DataSet.txt - Notepad
File Edit Format View Help
SU2

HDFS Architecture

putty cmd

add new directory

Movielens dataset

Hadoop fs -ls

MapReduce

Key value pair mapping

shuffle and sort

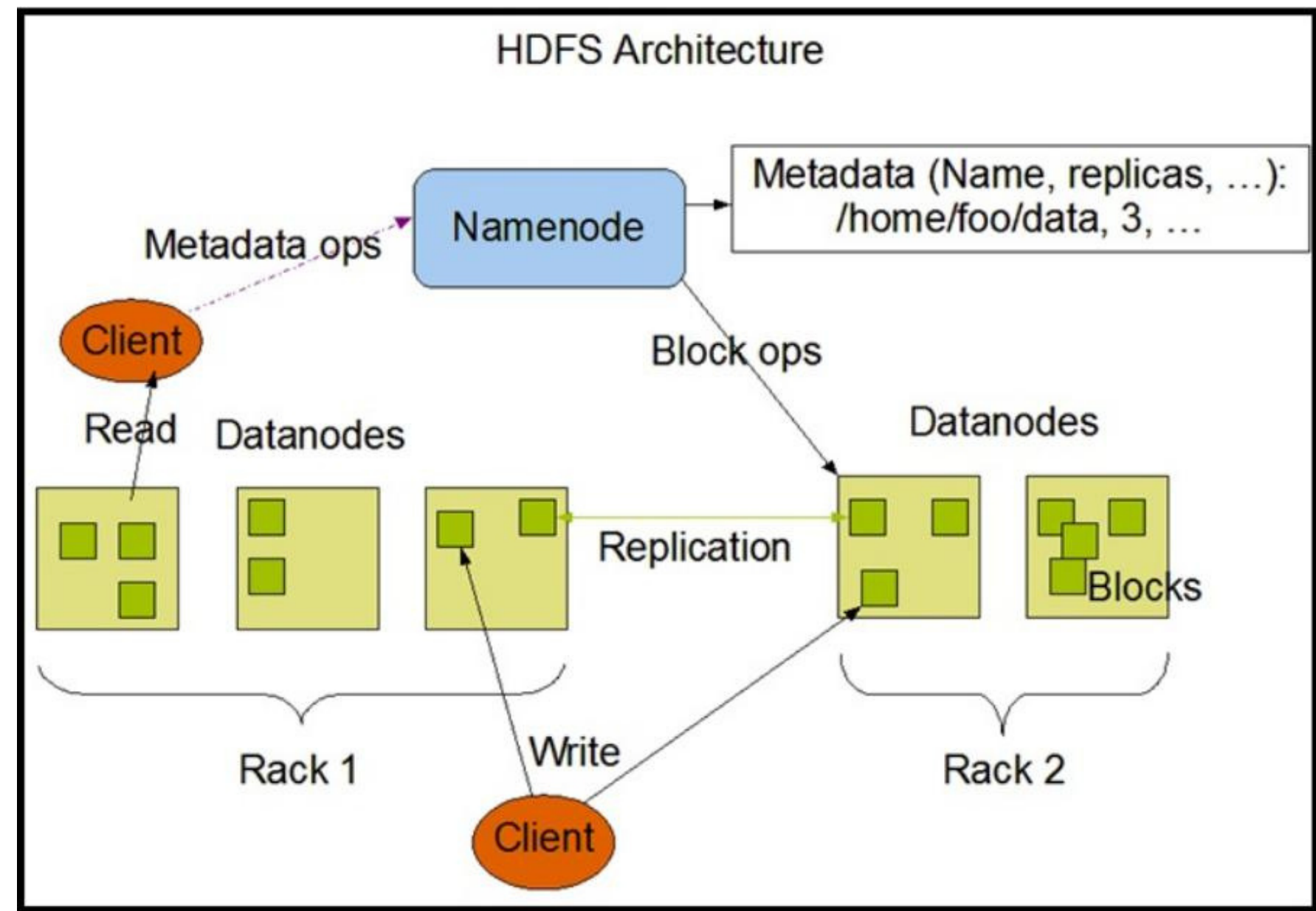
STREAMING MapReduce

Failures

18. video explain python
```

Ln 3, Col 18 100% Windows (CRLF) UTF-8

# HDFS Architecture



HDFS Architecture – main "NameNode" and multiple "data nodes" on a commodity hardware cluster. Data is then broken down into separate "blocks" that are distributed among the various data nodes for storage.



# hadoop fs -ls Command

---



## hadoop fs -ls

For a directory, it returns the list of files and directories whereas, for a file, it returns the statistics on the file.



## hadoop fs -lsr

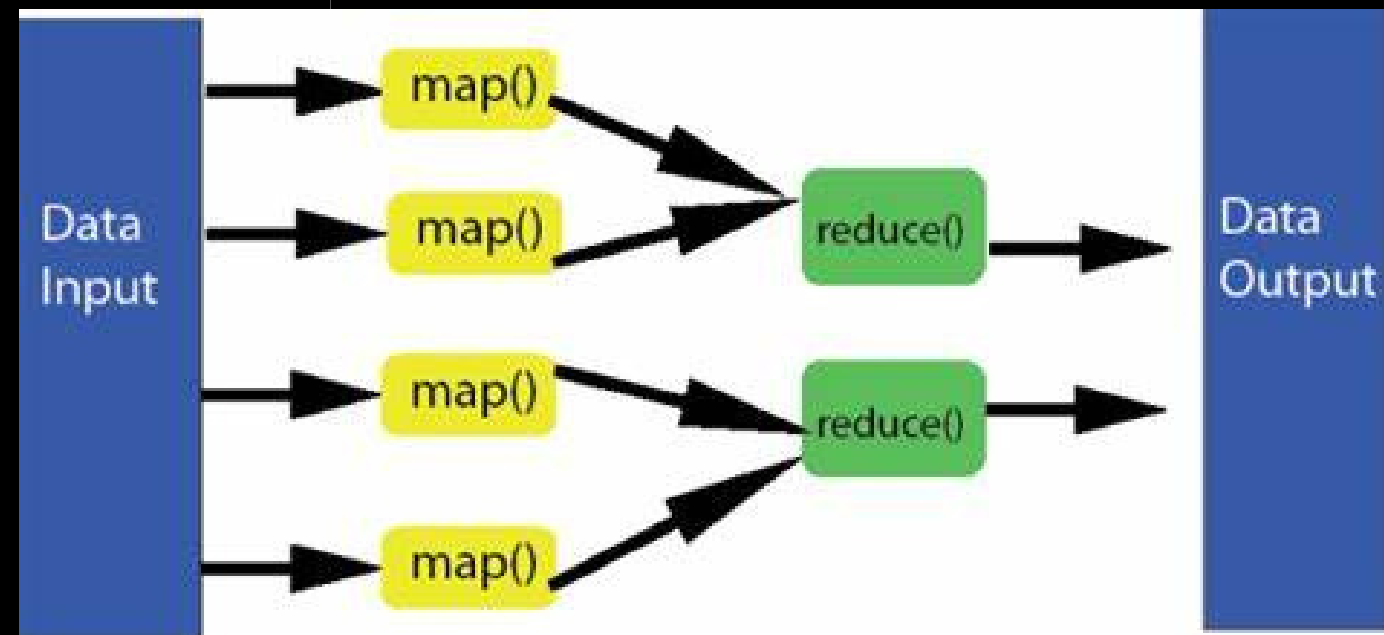
This is for recursively listing the directories and files under specific folders.



## Example

- Hadoop fs -ls
- Hadoop fs -lsr

# MapReduce



## DEFINITION

MapReduce is a programming model for writing applications that can process Big Data in parallel, on multiple nodes. MapReduce provides analytical capabilities for analyzing huge volumes of complex data.

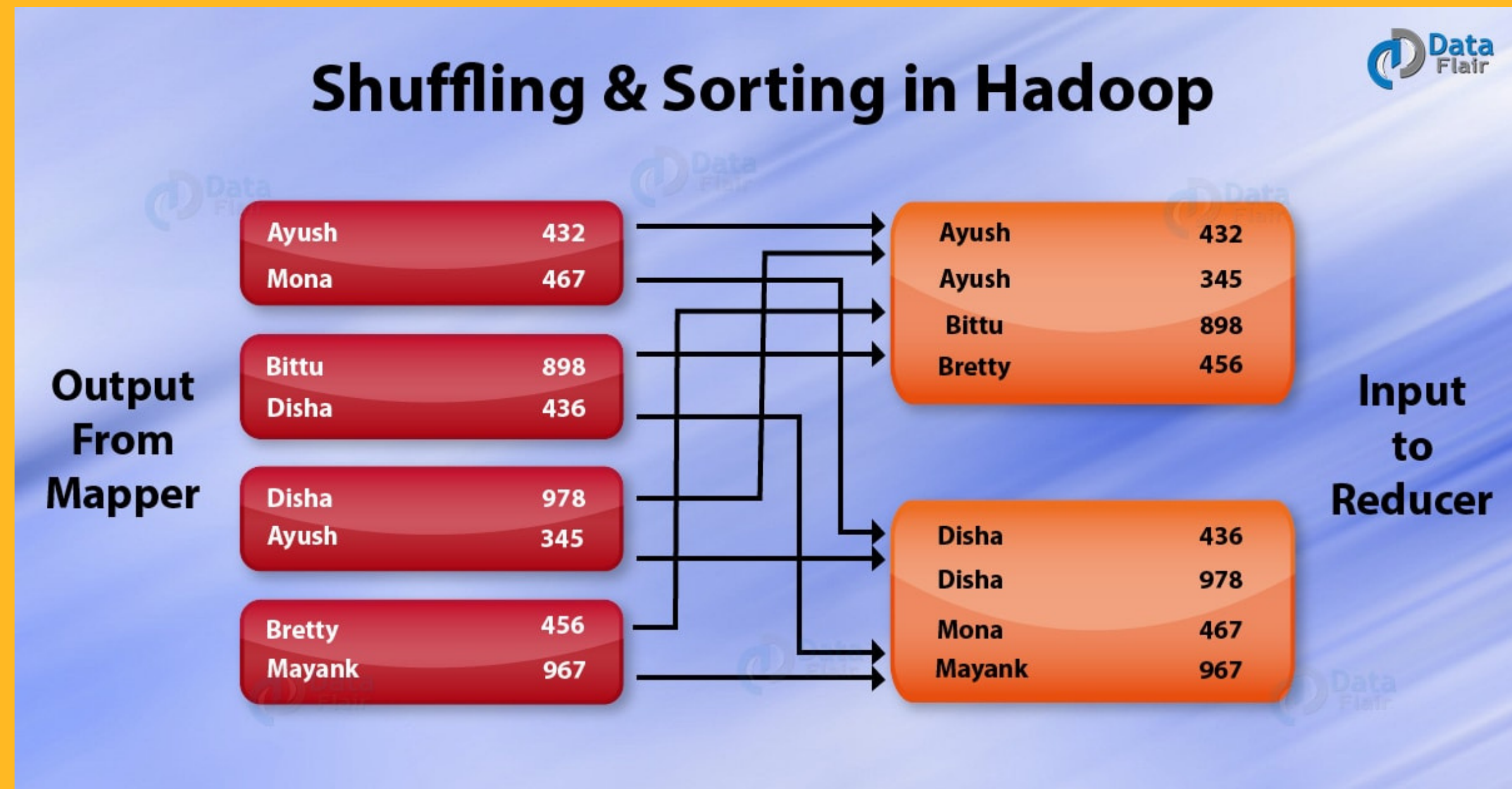
# Key-Value Pair Mapping

Is the process of binding the key to its associated value.  
For example, if a key of 'firstName' is associated with a value 'Bugs', it means that the array maps 'Bugs' to key firstName.

## EXAMPLE

key	value
firstName	Bugs
lastName	Bunny
location	Earth

# Shuffle and Sort



## How it works

The shuffle phase in Hadoop transfers the map output from Mapper to a Reducer in MapReduce. The sort phase in MapReduce covers the merging and sorting of map outputs. Data from the mapper are grouped by the key, split among reducers, and sorted by the key.



```
from mrjob.job import MRJob
from mrjob.step import MRStep

class RatingsBreakdown(MRJob):
    def steps(self):
        return [
            MRStep mapper=self.mapper_get_ratings,
                  reducer=self.reducer_count_ratings)
        ]

    def mapper_get_ratings(self, _, line):
        (userID, movieID, rating, timestamp) = line.split('\t')
        yield rating, 1

    def reducer_count_ratings(self, key, values):
        yield key, sum(values)

if __name__ == '__main__':
    RatingsBreakdown.run()
```

## My understanding

- MapReduce
- Getting the ratings
- Counting the ratings

The output was a list with the count of each movie.





# Thank you

---

**Enrico Dreyer**

31210783

---